Spatial Statistics 20 (2017) 110–124

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta

# Spatial clustering of curves with an application of satellite data

Gaetan Carlo, Girardi Paolo *, Pastres Roberto

*Dipartimento di Scienze Ambientali, Informatica e Statistica, Università Ca' Foscari di Venezia, Venezia, Italy*

ARTICLE INFO

*Article history:*
Received 20 September 2016
Accepted 24 January 2017
Available online 8 March 2017

*Keywords:*
Functional clustering
Markov random fields
Regression splines
Restoration–Maximization algorithm

ABSTRACT

Water quality indicators are important to identify risks to the environment, society and human health. The European Community Water Framework Directive establishes guidelines for the classification of all water bodies across Europe and chemical and biological indicators were used to this scope. In particular, the Chlorophyll type A index (Chl-a) is a shared indicator of trophic status and monitoring activities may be useful to explain its spatial distribution and to discover local dangerous behaviours (for example the anoxic events). Differently by the classical approach based on an "average" values over a period, we propose a functional clustering model that takes into account temporal and spatial dependence of Chl-a concentrations in the Adriatic Sea for defining appropriate clusters of sites. We use satellite monthly data, during the period 2002–2012, and we model the spatial dependence among the sites by means of a Markov random field model. Compared to similar attempts in literature by Jiang and Serban (2012) our formulation includes spatial covariates. This inclusion allows for more flexibility to obtain more homogeneous and representative clusters of sites in the Adriatic Sea. The estimation of the model and the identification of the number of clusters are carried out using a pseudolikelihood function. A small simulation study complements the real data analysis.



* Correspondence to: Dipartimento di Scienze Ambientali, Informatica e Statistica, Università Ca' Foscari - Venezia, Campus Scientifico, Via Torino, 155, I-30170 Mestre-Venezia, Italy.
*E-mail address:* paolo.girardi@unive.it (G. Paolo).

http://dx.doi.org/10.1016/j.spasta.2017.01.006
2211-6753/

## 1. Introduction

According to the European Water Framework Directive (WFD; 2000/60/EC), water bodies have to be monitored in order to achieve a "good ecological status" by 2015. Since the status of most European water bodies is affected by human activities, a regular monitoring activity includes the evaluation of the ecological status of each water body (such as bio-geochemical and hydromorphological parameters) with the scope to protect the environment, society and human health. A deterioration in water quality (i.e. eutrophication and cyanobacterial blooms) presents substantial risk and can have detrimental effects on the local economy.

In the sea water, the Chlorophyll type A indicator (Chl-a) is used as biomass indicator of primary producers (e.g. photosynthetic algae, from those unicellular to multicellular ones) in the water. The Chl-a level also increases as in eutrophic conditions, i.e. in presence of high concentration of nutrients and light availability. In such circumstance, marked algal blooms may be followed by nutrient depletion and a rapid decrease in algal biomass. The subsequent degradation may then lead to hypoxic or even anoxic events. The study related to the Chl-a concentrations and its trend may be useful to discover areas with different trophic status. This is particularly true for the Adriatic Sea (Marini et al., 2010; Giani et al., 2012). Time series of data can be easily gathered by remote satellite sensing techniques. However, few studies have classified areas on the basis of the entire temporal pattern of the time series. A common practice is to consider classification methods that compare only average values in a prefixed period (e.g. a month or a year). This limitation clearly leads to lose of information about the temporal pattern of the observed parameter.

Recent advances in environmental statistics provide new clustering methods based on functional data analysis (FDA) (Ramsay and Silverman, 2005). In FDA each time series is viewed as observations of a continuous function collected at a finite series of time points. In this setting, observations are functions and the fundamental unit of interest is the entire function or curve constructed from the observations collected over time. This approach has been applied in several environmental contexts: to classify time series concerning air quality indicators on sites as part of a monitoring network (Ignaccolo et al., 2008) or to regroup time series on water quality indicators (Pastres et al., 2011; Haggarty et al., 2012). Although with different approaches or formulations, these applications share the feature of grouping different sites together only when the observed time series have some common features, preserving sample information about the temporal pattern. However, they did not consider the spatial dependence in the clustering process and were limited to few sites.

In this respect a notable exception are two recent papers (Haggarty et al., 2015; Gaetan et al., 2016) where spatial dependence is accounted for by adjusting the $L_2$ dissimilarity measures calculated between two curves by means of a functional variogram (Giraldo et al., 2012). Another possibility still unexplored for environmental data is given by the bagging Voronoi classifiers (Secchi et al., 2013).

In this paper we follow a model based functional clustering approach (James and Sugar, 2003; Pastres et al., 2011; Haggarty et al., 2012, 2015) in which time series are supposed to be generated by a mixture of latent distributions. Because, in our motivating example, time series are collected on a regular grid or network we take into account the spatial dependence using a Markov random model as mixing distribution (Jiang and Serban, 2012). In addition we extend the modelling proposal of that by including different level of spatial dependence in the latent process, namely in the conditional mean and in the interacting potentials. This inclusion, motivated by our real problem, leads to more spatially homogeneous and consistent clustering of the water bodies in the Adriatic Sea based on Chl-a concentrations according to the current knowledge on the Chl-a dynamics in that zones.

The structure of the paper is the following. In Section 2 we highlight the main features of the data that we used. In Section 3 we introduce our model proposal and a small simulation study is presented in Section 4 for demonstrating the clustering performances. Section 5 reports our finding for the classification. A brief discussion about the results (Section 6) ends the paper.

## 2. Chl-a concentrations over the Adriatic Sea

Among the European waters, the Adriatic Sea reports some singular characteristics: it's almost land-locked basin separated from the central Mediterranean by the Strait of Otranto; although a fairly